1W-8

ユーザビリティの高いGUIベースの決定木学習ツール“ID3E”の開発

並木　翼†　　菊池　浩明‡
東海大学電子情報学部†‡

1．はじめに

近年,膨大なデータの中から有益な情報のみを抽出する技術,データマイニングが注目されている．本研究ではデータマイニング手法の一つである決定木学習に焦点を当てる．

決定木学習が可能なデータマイニングツール[1][2][3]は数多く存在する．しかし,データマイニングの専門知識が不可欠であったり,コマンドライン上で実行するCUIツールが多い，といった理由から初心者に敬遠されやすい．そこで,本研究では,誰にでも扱いやすいGUI決定木ツールの開発を目的とする．また,開発した決定木学習ツールの有用性を確かめるために，従来の決定木ツールとのユーザビリティに関する比較実験の結果を報告する．

2．決定木学習ツール“ID3E”の開発

2.1．従来ツールの問題点

2.1.1．データフォーマットの特殊性[1][2][3]

決定木学習ツールのほとんどが,そのツール独自の学習データフォーマットを採用している．例えば，C4.5R8[3]は属性名宣言とデータファイルを二つのファイルに分けて記述しなければならず，他のツールとの互換性を損なっていた．

2.1.2．全体把握の困難性[2][3]

コマンドベースツールの決定木はテキストで表現される．１ノードにつき１行のテキストで表されるので，決定木が大きくなればなるほど改行の個数が多くなり，スクリーンに収まらなくなる．決定木のバランスが一目で分かりづらい．

2.1.3．枝刈り(チューニング)の不敏性[1][2][3]

従来ツールでは枝刈りをするたび，パラメータを変えてコマンドを再入力しなければならない．図１に示されるように，この際入力と木の再構築は冗長である．

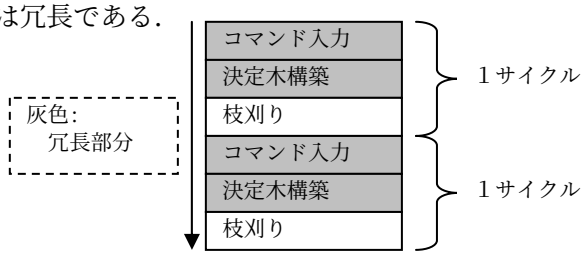

図１．枝刈りの不敏性

2.2．開発方針

2.2.1．概要

データマイニングの初心者でも判りやすく使いやすい決定木学習ツールを目指して開発を進める．そのためには,面倒な操作を省いた扱いやすいインターフェイス,視覚的にわかりやすいグラフィカルなアウトプットに重点を置く必要がある．

2.2.2．　学習データ記述の簡略化

データは全てCSV形式とする．これにより，表計算アプリケーションと連携できるようにした．

2.2.3．決定木Viewer

ユーザビリティの向上を図るために決定木を視覚化する．本ツールの基本機能と連動しており,枝刈りや決定木評価が行われると結果が動的に反映される．スクリーンのスライド,拡大縮小などの操作は全てマウスで行う（図2参照）．

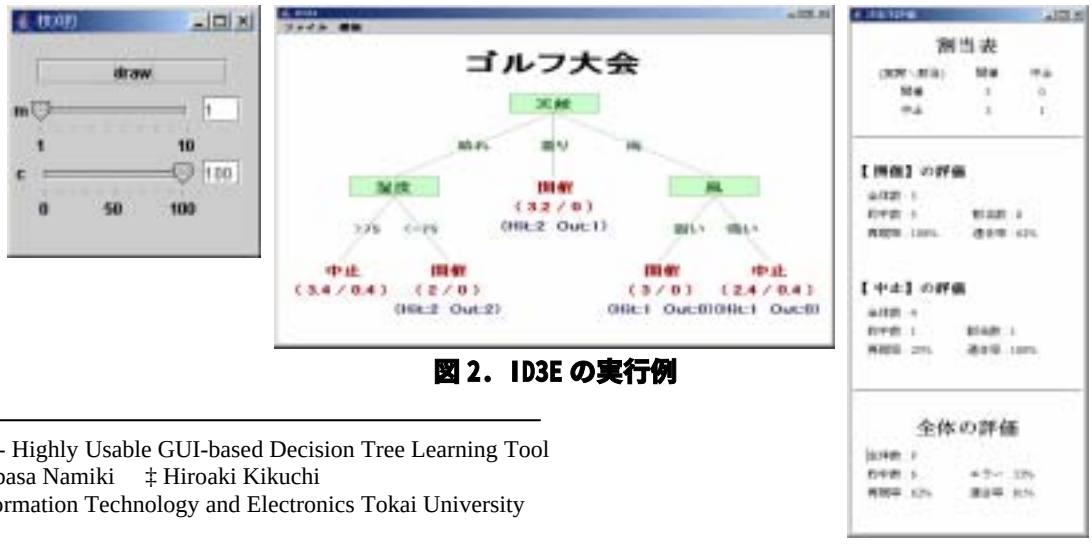

図２．ID3Eの実行例

ID3E - Highly Usable GUI-based Decision Tree Learning Tool
† Tsubasa Namiki　‡ Hiroaki Kikuchi
Information Technology and Electronics Tokai University

2.2.4. 枝刈り(チューニング)の改善と補助機能

枝刈りによる冗長な決定木の再構築を防ぐために,元の木をメモリ上に常駐させておく.更に,枝刈りパラメータをスライドバーで指定できるようにしてユーザビリティを高める（図2左参照）.枝刈りの木に対する評価が動的に反映され,枝刈り木の効果を判別しやすくしている.

2.2.5. 決定木評価機能

決定木のエラー率に加えて,各クラスの再現率・適合率を表示する機能を実装した（図2右参照）.また，決定木Viewerの木の葉に評価データについての正当数(Hit,Out)を表示している.

2.3. 実装について

ID3Eの実装にはJava言語を用いた．表１に実装した主要機能の一部を示す.

表1. ID3Eの主要機能一覧表

決定木構築機能	
欠損値対応	構築用データの虫食いを補完(１票を分割)する機能
連続属性対応	1,2,3,4,5 などの数値の属性に対応
決定木枝刈り機能	
重み付け枝刈り(M)	指定票数 M を達していないノードから下の枝を刈る
信頼度枝刈り(CF)	指定信頼度 CF を達していないノードから下の枝を刈る
決定木イメージ機能	
決定木 Viewer	決定木をスライド拡大縮小表示するための Viewer
決定木評価機能	
データ評価の指定	学習材の木に評価データを自動割当して，精度を評価
適合・再現率	割当表を用いて，クラスごとに適合・再現率を求める
決定木総合評価	決定木全体のエラー率・的中数などを求める

3. ユーザビリティ評価実験

3.1. 概要

開発したツールのユーザビリティの向上を確かめるために,ある作業にかかった時間を従来のツールと開発したツールとで比較をする実験を行う．3.3 節の作業について，設問を解くまでにかかった時間を各ツールで計測した．設問の解答ミスも評価基準に入れる．各作業によってツールの利用方法は異なる．なお,比較対象ツールには，C4.5R8[3]を使用する.

3.2. 実験環境

実験マニュアルと結果記入用紙を当研究室学生(データマイニング未経験者)に配布して，2004 年 11 月 22 日からの７日間に実験した．被験者は合計７人で，内５人は当大学のコンピュータ室に集まってもらい，実験責任者の指示のもとに作業を行った.

3.3. 作業内容

次の３つの作業を行った．(3)の一部を図３に示す.

(1) 決定木評価 – ツール全体の作業効率を計る
(2) 手動割当評価 – 木の目視探索効率を計る
(3) 枝刈箇所特定 – 木全体の把握効率を計る

(3). 枝刈箇所特定
1. タイムウォッチスタート
2. ID3G.jar を実行
3. ウィンドウに UTest.csv を Drag&Drop
4. 右クリックで枝刈りウィンドウを出す
5. m の値を変えて,刈られた箇所を特定する
6. 刈られた箇所を用紙に記入
7. 用紙の空欄が埋まるまで 5,6 を繰り返す
8. タイムウォッチストップ
9. かかった時間を用紙に記入

図 3. 実験マニュアル抜粋

3.4. 実験結果

表 2. 計 7 人の実験結果集計

作業	提案ツール“ID3E”			従来ツール“C4.5R8”		
	平均[s]	SD	総ミス	平均[s]	SD	総ミス
(1)	365	261	0	282	70	0
(2)	297	123	3	494	338	3
(3)	373	166	3	440	141	6

3.5. 考察

ID3E の総合作業時間が従来ツールの 85%で済むことがわかった．決定木そのものを扱う作業(2)(3)では従来ツールの方が時間がかかっており，ミスも多い．決定木をグラフィカルに可視化したことによる恩恵であるといえる．しかし，作業(1)は従来ツールには及ばなかった．原因はインターフェイスの不備,GUI グラフィカル処理による遅延と考えられる．これらを今後の課題とする.

4. おわりに

ユーザビリティの高い GUI ベースの決定木学習ツールを開発した．また,ユーザビリティ評価実験により新ツールの総合作業時間が従来ツール(C4.5R8[3])の 85%で済むことがわかった．今後の課題はインターフェイスの更なる向上である.

参考文献

[1] The University of Waikato, “Weka-3.4.2”, 2004年5月.
[2] MUSASHIプロジェクト, “MUSASHI-1.0.3FC2”, 2004年5月.
[3] J. Ross Quinlan, “C4.5: Programs for Machine Leaning”, Morgan Kaufmann pub., 1993.